Biometrics 66, 1096–1106
December 2010
DOI: 10.1111/j.1541-0420.2010.01395.x

Bias-Corrected Diagonal Discriminant Rules for High-Dimensional Classification

Song Huang,[1,*] Tiejun Tong,[2,] and Hongyu Zhao[3,4,***]**

[1]Program of Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, U.S.A.
[2]Department of Applied Mathematics, University of Colorado, Boulder, Colorado 80309, U.S.A.
[3]Department of Epidemiology and Public Health, Yale University, New Haven, Connecticut 06520, U.S.A.
[4]Department of Genetics, Yale University, New Haven, Connecticut 06520, U.S.A.
[*]*email:* song.huang@yale.edu
[**]*email:* tiejun.tong@colorado.edu
[***]*email:* hongyu.zhao@yale.edu

Summary. Diagonal discriminant rules have been successfully used for high-dimensional classification problems, but suffer from the serious drawback of biased discriminant scores. In this article, we propose improved diagonal discriminant rules with bias-corrected discriminant scores for high-dimensional classification. We show that the proposed discriminant scores dominate the standard ones under the quadratic loss function. Analytical results on why the bias-corrected rules can potentially improve the predication accuracy are also provided. Finally, we demonstrate the improvement of the proposed rules over the original ones through extensive simulation studies and real case studies.

Key words: Bias correction; Diagonal discriminant analysis; Discriminant score; Large p small n; Tumor classification.

1. Introduction

Class prediction using high-dimensional data such as microarrays has been recognized as an important problem since the seminal work of Golub et al. (1999). A variety of methods have been developed and compared, including discriminant analysis and its extensions (Dudoit, Fridlyand, and Speed, 2002; Ghosh, 2003; Zhu and Hastie, 2004; Huang and Zheng, 2006; Shen et al., 2006; Wu, 2006; Guo, Hastie, and Tibshirani, 2007; Pang, Tong, and Zhao, 2009), random forests (Breiman, 2001; Statnikov, Wang, and Aliferis, 2008), support vector machines (Furey et al., 2000; Lee, Lin, and Wahba, 2004; Vapnik and Kotz, 2006), dimension reduction methods (Antoniadis, Lambert-Lacroix, and Leblanc, 2003; Dai, Lieu, and Rocke, 2006), and nearest shrunken centroids methods (Tibshirani et al., 2002, 2003; Wang and Zhu, 2007; Dabney and Storey, 2007). Also see review papers with extensive comparison studies by Dudoit et al. (2002), Lee et al. (2005), and Statnikov et al. (2008).

In high-dimensional microarray data classification, it is common that the number of training samples, n, is much smaller than the number of features examined, p. This "large p small n" paradigm has posed numerous statistical challenges to most classical classification methods, such as the well-known linear discriminant analysis (LDA) and the quadratic discriminant analysis (QDA), because the sample covariance matrices are singular. This greatly limits the usage of both methods in high-dimensional data classification. To overcome the singularity problem, various approaches that rely on a diagonal approximation to the covariance matrices have been proposed. This leads to the so-called diagonal discriminant rules, which have been widely used for high-dimensional data (Dudoit et al., 2002; Speed, 2003; Tibshirani et al., 2003; Dettling, 2004; Ye et al., 2004; Dabney, 2005; Lee et al., 2005; Pique-Regi, Ortega, and Asgharzadeh, 2005; Asyali et al., 2006; Noushath, Kumar, and Shivakumara, 2006; Shieh, Jiang, and Shih, 2006; Wang and Zhu, 2007; Natowicz et al., 2008; Pang et al., 2009). In practice, the most commonly used diagonal discriminant rules for high-dimensional data are the diagonal LDA (DLDA) and the diagonal QDA (DQDA) rules introduced by Dudoit et al. (2002). Due to the relatively small n, the diagonal discriminant rules, which ignore the correlation among features, performed remarkably well compared with the more sophisticated methods in terms of both accuracy and stability (Dudoit et al., 2002; Dettling, 2004; Lee et al., 2005; Pang et al., 2009). In addition, DLDA and DQDA are easy to implement and not very sensitive to the number of predictor variables (Dudoit et al., 2002). Bickel and Levina (2004) conducted a theoretical study of this phenomenon and proved that diagonal discriminant rules can indeed outperform Fisher's LDA when $p > n$.

The diagonal discriminant rules have been shown to perform well for high-dimensional data with small sample sizes, but suffer from the serious drawback of biased discriminant scores. In this article, we propose to correct the biases in the discriminant scores of diagonal discriminant analysis. Before we proceed, it is worth pointing out that the idea of bias correction in discriminant analysis is not entirely new (Ghurye and Own, 1969; Moran and Murphy, 1979; McLachlan, 1992). For instance, Moran and Murphy (1979) proposed several bias correction methods for the plug-in discriminant scores under

the condition that the sample size for each class, n_k, is larger than p. However, the improvement of their bias-corrected rules is not significant (James, 1985; McLachlan, 1992), mainly because the dominant term of the bias, p/n_k, is not large. This has, at least partially, discouraged the popularity of the previously proposed bias-corrected discriminant rules. For microarray data, however, the ratio p/n_k can be very large. As a consequence, the commonly used discriminant rules, for example, DQDA and DLDA, may result in low prediction accuracy, especially when the design is fairly unbalanced.

The remainder of the article is organized as follows. In Section 2, we introduce the notation and briefly review the diagonal discriminant rules. In Section 3, we derive the bias-corrected estimators of the discriminant scores and show that they dominate the original ones. In Section 4, we present some analytical results on why the bias-corrected rules can potentially increase the overall prediction accuracy. We then conduct extensive simulation studies to investigate the performance of the proposed methods in Section 5, and apply them to three real microarray data sets in Section 6. Finally, we conclude the paper in Section 7 with discussions and future directions.

2. Diagonal Discriminant Analysis

Suppose we have K distinct classes and samples from each class follow a p-dimensional multivariate normal distribution with mean vector μ_k and covariance matrix Σ_k, where $k = 1, \ldots, K$. Assume we observe n_k i.i.d. random samples from the kth class, that is,

$$\boldsymbol{x}_{k,1}, \ldots, \boldsymbol{x}_{k,n_k} \overset{i.i.d.}{\sim} MVN(\boldsymbol{\mu}_k, \Sigma_k).$$

The total sample size is then $n = \sum_{k=1}^{K} n_k$. The principal goal of discriminant analysis is to predict the class label for a new observation, $\boldsymbol{y}$. Let π_k denote the prior probability of observing a sample from the kth class with $\sum_{k=1}^{K} \pi_k = 1$. The QDA decision rule is to assign $\boldsymbol{y}$ to class $\arg\min_k d_k^Q(\boldsymbol{y})$, where $d_k^Q(\boldsymbol{y})$ is the discriminant score defined as in Friedman (1989), that is,

$$d_k^Q(\boldsymbol{y}) = (\boldsymbol{y} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_k) + \ln|\Sigma_k| - 2\ln\pi_k.$$

Minimizing $d_k^Q(\boldsymbol{y})$ over k is equivalent to maximizing the corresponding posterior probabilities.

In practice, the population parameters of the multivariate normal distributions are unknown and usually are estimated from the training data set, with $\boldsymbol{\mu}_k$ by the sample means, $\hat{\boldsymbol{\mu}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} \boldsymbol{x}_{k,i}$, and Σ_k by the sample covariance matrices, $\hat{\Sigma}_k = \frac{1}{n_k - 1}\sum_{i=1}^{n_k} (\boldsymbol{x}_{k,i} - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_{k,i} - \hat{\boldsymbol{\mu}}_k)^T$. In addition, the prior probability π_k is commonly estimated by n_k/n and treated as a constant in classification problems (Friedman, 1989; Guo et al., 2007). The above estimates of parameters lead to the following sample version of $d_k^Q(\boldsymbol{y})$:

$$\hat{d}_k^Q(\boldsymbol{y}) = (\boldsymbol{y} - \hat{\boldsymbol{\mu}}_k)^T \hat{\Sigma}_k^{-1} (\boldsymbol{y} - \hat{\boldsymbol{\mu}}_k) + \ln|\hat{\Sigma}_k| - 2\ln\pi_k. \quad (1)$$

One important special case of QDA is to assume that the covariance matrices are all the same, that is, $\Sigma_k = \Sigma$ for all k. This leads to LDA, with the simplified discriminant score given by

$$d_k^L(\boldsymbol{y}) = (\boldsymbol{y} - \boldsymbol{\mu}_k)^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_k) - 2\ln\pi_k.$$

The corresponding sample version of $d_k^L(\boldsymbol{y})$ is then

$$\hat{d}_k^L(\boldsymbol{y}) = (\boldsymbol{y} - \hat{\boldsymbol{\mu}}_k)^T \hat{\Sigma}^{-1} (\boldsymbol{y} - \hat{\boldsymbol{\mu}}_k) - 2\ln\pi_k, \quad (2)$$

with the pooled sample covariance matrix estimate $\hat{\Sigma} = \frac{1}{n-K}\sum_{k=1}^{K}(n_k - 1)\hat{\Sigma}_k$.

QDA and LDA are expected to perform well if the multivariate normal assumption is satisfied and good "plug-in" estimates of the population parameters are available (Friedman, 1989). In general, LDA is more popular than QDA, largely due to its simplicity and robustness to the violations of the underlying distribution assumption and the common covariance matrices assumption (James, 1985). To make LDA work, we require that $n \geqslant p$ to ensure the nonsingularity of $\hat{\Sigma}$. Similarly for QDA, we require that $n_k \geqslant p$ for each class.

When p is greater than n, we may regularize the covariance matrix estimates with generalized matrix inverse or shrinkage to address the singularity problem. However, these estimators are usually unstable due to the limited number of observations (Guo et al., 2007). In 2002, Dudoit et al. proposed to use DQDA and DLDA for classifying tumors using microarray data. Specifically, they assumed the covariance matrices to be diagonal by replacing the off-diagonal elements of $\hat{\Sigma}_k$ or $\hat{\Sigma}$ with zeros. For DQDA, we have $\hat{\Sigma}_k = \mathrm{diag}(\hat{\sigma}_{k1}^2, \ldots, \hat{\sigma}_{kp}^2)$, which simplifies equation (1) to

$$\hat{d}_k^Q(\boldsymbol{y}) = \sum_{i=1}^{p} (y_i - \hat{\mu}_{ki})^2/\hat{\sigma}_{ki}^2 + \sum_{i=1}^{p} \ln\hat{\sigma}_{ki}^2 - 2\ln\pi_k. \quad (3)$$

For DLDA, we have $\hat{\Sigma} = \mathrm{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2)$, which simplifies equation (2) to

$$\hat{d}_k^L(\boldsymbol{y}) = \sum_{i=1}^{p} (y_i - \hat{\mu}_{ki})^2/\hat{\sigma}_i^2 - 2\ln\pi_k. \quad (4)$$

3. Bias-Corrected Diagonal Discriminant Analysis

In this section, we first show that $\hat{d}_k^Q(\boldsymbol{y})$ and $\hat{d}_k^L(\boldsymbol{y})$ are biased. We then propose several bias-corrected estimators for the discriminant scores and demonstrate their superiority over the original ones. Denote equation (3) as

$$\hat{d}_k^Q(\boldsymbol{y}) = \hat{L}_{k1} + \hat{L}_{k2} - 2\ln\pi_k,$$

where $\hat{L}_{k1} = \sum_{i=1}^{p}(y_i - \hat{\mu}_{ki})^2/\hat{\sigma}_{ki}^2$ and $\hat{L}_{k2} = \sum_{i=1}^{p}\ln\hat{\sigma}_{ki}^2$. Denote the true discriminant score as

$$d_k^Q(\boldsymbol{y}) = L_{k1} + L_{k2} - 2\ln\pi_k,$$

where $L_{k1} = \sum_{i=1}^{p}(y_i - \mu_{ki})^2/\sigma_{ki}^2$ and $L_{k2} = \sum_{i=1}^{p}\ln\sigma_{ki}^2$. In Web Appendix A, we show that the following two estimators are *unbiased* for L_{k1} and L_{k2}, respectively,

$$\tilde{L}_{k1} = \frac{n_k - 3}{n_k - 1}\hat{L}_{k1} - \frac{p}{n_k},$$
$$\tilde{L}_{k2} = \hat{L}_{k2} - p\left\{\Psi\left(\frac{n_k - 1}{2}\right) - \ln\left(\frac{n_k - 1}{2}\right)\right\},$$

where $\Psi(\cdot)$ is the digamma function (Abramowitz and Stegun, 1972). Based on the above two unbiased estimators, we define

$$\tilde{d}_k^Q(\boldsymbol{y}) = \tilde{L}_{k1} + \tilde{L}_{k2} - 2\ln\pi_k,$$

which is a bias-corrected discriminant score of DQDA. We refer to the corresponding rule as the bias-corrected DQDA (BQDA).

For DLDA, we denote equation (4) as $\hat{d}_k^L(\boldsymbol{y}) = \hat{L}_k - 2\ln\pi_k$ and the corresponding true discriminant score as $d_k^L(\boldsymbol{y}) = L_k - 2\ln\pi_k$, where $\hat{L}_k = \sum_{i=1}^p (y_i - \hat{\mu}_{ki})^2/\hat{\sigma}_i^2$ and $L_k = \sum_{i=1}^p (y_i - \mu_{ki})^2/\sigma_i^2$. Also in Web Appendix A, we show that the following estimator is *unbiased* for L_k:

$$\tilde{L}_k = \frac{n-K-2}{n-K}\hat{L}_k - \frac{p}{n_k},$$

which leads to the bias-corrected DLDA (BLDA) with

$$\tilde{d}_k^L(\boldsymbol{y}) = \tilde{L}_k - 2\ln\pi_k.$$

Further, we have

THEOREM 1. *Under the quadratic loss function, we have*

(i) the discriminant score of BQDA, $\tilde{d}_k^Q$, dominates the discriminant score of DQDA, $\hat{d}_k^Q$, when $n_k > 5$; and

(ii) the discriminant score of BLDA, $\tilde{d}_k^L$, dominates the discriminant score of DLDA, $\hat{d}_k^L$, when $n > K + 4$.

The proof of Theorem 1 is shown in Appendix A. The maximum likelihood estimators (MLE), $\hat{\sigma}_{ki,ML}^2 = \frac{n_k-1}{n_k}\hat{\sigma}_{ki}^2$, are also common for estimating σ_{ki}^2 (Guo et al., 2007). By plugging $\hat{\sigma}_{ki,ML}^2$ into equation (3), we obtain the discriminant score of MLE-based DQDA (MQDA). In practice, there is usually no clear indication between $\hat{\sigma}_{ki,ML}^2$ and $\hat{\sigma}_{ki}^2$, as to which estimator performs better when n is small. It is worth pointing out that, when the bias correction technique is applied, DQDA and MQDA lead to the same discriminant score so that we do not need to distinguish the two methods any more. A similar result can be established for the MLE-based DLDA (MLDA).

4. Prediction Accuracy

In this section, we compare the performance of the bias-corrected discriminant rules with that of the original ones. The prediction accuracy is a common measure for evaluating the performance of a discriminant rule. It is defined as the proportion of samples classified correctly in the test set and is usually used for a balanced experimental design (Dudoit et al., 2002). However, when the design is unbalanced, a classification method favoring the majority class may have a high prediction accuracy (Qiao and Liu, 2009). There are many evaluation criteria for unbalanced designs, for example, *G*-mean, *F*-measure, recall, class-weighted accuracy (CWA), among others (Chen, Liaw, and Breiman, 2004; Cohen et al., 2006; Qiao and Liu, 2009). All of the above performance metrics can be viewed as functions of the classification matrix formed by the probabilities $Pr(\text{True Class} = i, \text{Predicted Class} = j)$. Each matrix has its own advantages and limitations (Chen et al., 2004; Cohen et al., 2006). In this article, we apply the CWA criterion (Cohen et al., 2006), which is defined as

$$CWA = \sum_{k=1}^K w_k a_k,$$

where a_k are the per-class predication accuracies and w_k are nonnegative weights with $\sum_{k=1}^K w_k = 1$. For simplicity, we assume equal weights, that is, $w_k = 1/K$, and set the prior probability $\pi_k = 1/K$ as well. Note that CWA is equivalent to one of the criteria proposed by Qiao and Liu (2009), which they referred to as the "mean within group error with one-step fixed weights" criterion.

In what follows, we establish some analytical results for the bias-corrected rules. For simplicity of exposition, we consider the binary classification ($K = 2$) with the following three assumptions:

(i) the variances are known and equal (without loss of generality, we assume that $\sigma_{ki}^2 = 1$);

(ii) $n_1 < n_2$, that is, the class 1 is the minority class and the class 2 is the majority class; and

(iii) the covariance matrix of the test data is diagonal.

Under the above assumptions, we have $\hat{d}_k^L = \sum_{i=1}^p (y_i - \hat{\mu}_{ki})^2$ for DLDA, and $\tilde{d}_k^L = \hat{d}_k^L - p/n_k$ for BLDA. Denote $\hat{D} = \hat{d}_1^L - \hat{d}_2^L$. For DLDA, we assign $\boldsymbol{y}$ to the minority class if $\hat{D} < 0$; otherwise, we assign it to the majority class. For BLDA, the decision boundary is $U = p(\frac{1}{n_1} - \frac{1}{n_2})$ instead of a usual zero. It is easy to see that the expected change of prediction accuracy caused by the bias correction, $Pr_{\hat{D},k}$, is given as

$$Pr_{\hat{D},k} = Pr(0 < \hat{D} < U | \boldsymbol{y} \in \text{class })k, k = 1, 2.$$

Note that for an unbalanced design, the prediction accuracy of the minority class always increases and that for the majority class always decreases because of the bias correction. The overall CWA change, Pr_Δ, is given as

$$Pr_\Delta = Pr_{\hat{D},1} - Pr_{\hat{D},2}, \tag{5}$$

where a positive Pr_Δ indicates an overall improvement on the classification performance.

By the Lindeberg condition of the central limit theorem (Lehmann, 1998), it can be shown that when $p \to \infty, \hat{D}$ converges in distribution to $N(-\delta + U, 4b_1\delta + c)$ if $\boldsymbol{y}$ is from class 1, and $\hat{D}$ converges in distribution to $N(\delta + U, 4b_2\delta + c)$ if $\boldsymbol{y}$ is from class 2, where $\delta = \sum_{i=1}^p (\mu_{1i} - \mu_{2i})^2, b_1 = 1 + 1/n_2, b_2 = 1 + 1/n_1$, and $c = 2p\{(2n_1 + 1)/n_1^2 + (2n_2 + 1)/n_2^2\}$. Note that δ is the squared Euclidean distance between two samples. Further, we have

THEOREM 2. *Under Assumptions (i)–(iii), the overall CWA change, Pr_Δ, is positive when $0 < \delta/p \leqslant 2$ and $p \to \infty$.*

The proof of Theorem 2 is shown in Appendix B. Theorem 2 suggests that the bias correction improves the overall predication accuracy as p goes large and δ is bounded by $2p$. It is also worth mentioning that the proposed decision boundary U is asymptotically optimal under certain situations. By the definition of Pr_Δ, it is easy to see that the optimal decision boundary, U_{opt}, can be achieved at the intersection of the two limiting normal distributions, $N(-\delta + U, 4b_1\delta + c)$ and $N(\delta + U, 4b_2\delta + c)$. When δ is not large and/or the sample sizes, n_1 and n_2, are at least moderately large, we have $4b_1\delta + c \approx 4b_2\delta + c$ and thus $U_{opt} \approx U$. In general, as $4b_1\delta + c < 4b_2\delta + c, U_{opt}$ is slightly larger than U when δ is close to zero, and vice versa when δ is large. Note that U_{opt} depends on the quantity of δ so it is unknown in practice. Simulation study (not shown) indicates that the discriminant rules based on U and an estimated $\hat{U}_{opt}$ perform similarly when an accurate estimate of the unknown δ can be

obtained. While if a less-accurate estimate of δ is employed, the performance of $\hat{U}_{opt}$ can be unsatisfactory. In addition, U_{opt} is obtained only in the asymptotic sense so it may not work well when p is small. For the above reasons, in what follows we will only focus on the decision boundary U but not $\hat{U}_{opt}$. Note that the cross-validation (CV) method can be used as an alternative to select the decision boundary. Simulation study (not shown) indicates that it performs similarly as U when the sample size of each class is large. While for a small n_1 or n_2, CV is unstable and consequently the performance of BLDA is not satisfactory, as indicated in Braga-Neto and Dougherty (2004), Fu et al. (2005), and Isaksson et al. (2008).

When p is small, the overall change of CWA is given as

$$Pr_{\Delta} = \int_{-\infty}^{+\infty} Pr_{\hat{D}|\boldsymbol{y}} f_1(\boldsymbol{y}) d\boldsymbol{y} - \int_{-\infty}^{+\infty} Pr_{\hat{D}|\boldsymbol{y}} f_2(\boldsymbol{y}) d\boldsymbol{y}, \quad (6)$$

where $f_t(\boldsymbol{y}) = \prod_{i=1}^{p} \phi(y_i, \mu_{ti}, \sigma_{ti}^2)$ is the joint density function, and $Pr_{\hat{D}|\boldsymbol{y}}$ is the expected change of prediction accuracy given an observation $\boldsymbol{y}$. One way to obtain Pr_{Δ} in equation (6) is to use the numerical integration approach as shown in Appendix C.

5. Simulation Studies

In this section, we conduct extensive simulation studies to assess the performance of different discriminant rules under various settings. We explain in detail both simulation designs and results of the simple binary classification, as well as some more complicated scenarios such as the multiple classification.

5.1 *Simulation Design*

We draw n_k training samples, $\boldsymbol{x}_{k,i}$, and m_k test samples, $\boldsymbol{y}_{k,j}$, from a G-dimensional multivariate normal distribution, $\boldsymbol{x}_{k,i}, \boldsymbol{y}_{k,j} \overset{i.i.d.}{\sim} MVN(\boldsymbol{\mu}_k, \Sigma_k)$, where $i = 1, \ldots, n_k$ and $j = 1, \ldots, m_k$. Usually G is large for microarray studies. For binary classification problem, we have $K = 2$. Note that we only choose p genes from all G genes for classification based on certain feature selection criteria.

We first evaluate CWA directly under different simulation settings with the assumptions stated in Section 4. We assume that all p genes are informative and the differences of the two group means are the same across all p genes. Note that if we increase p, the overall strength of the signal, $\delta = \sum_{i=1}^{p} (\mu_{1i} - \mu_{2i})^2$, becomes stronger and eventually both the bias-corrected methods and the biased methods will classify samples with 100% accuracy. To visualize the comparison results for different p values, we fix δ as a constant. For large p, we compute the change of CWA directly from equation (5); otherwise, CWA is approximated by integrating equation (6) numerically. In both cases, we assume genes are independent from each other with variances equal to one. When the genes are dependent with unknown variances, we go through the regular classification procedure to estimate the prediction accuracy as outlined below.

Next, we consider simulation settings that are closer to real data structures where genes are correlated to each other. We set the first g genes are informative, for example, $\mu_{1i} = 0.5$ and $\mu_{2i} = 0, i = 1, \ldots, g$, and the rest of $(G - g)$ genes have $\mu_{1i} = \mu_{2i} = 0, i = g + 1, \ldots, G$. Note that no feature selection procedure is involved here yet. We select the first p genes for classification. If $p \leqslant g$, all p genes are informative. If $p > g$, all of the g informative genes and $(p - g)$ noninformative genes are selected. Usually we let $g \ll G$ due to the fact that most of the genes are not differentially expressed in microarray experiments, for example, $G = 10{,}000$ and $g = 50$. Similarly as in Guo et al. (2007), we use block diagonal correlation structures to model the dependence among genes. Specifically, we partition the G genes into H equal-sized blocks with $H = G/g$. We have

$$\Sigma_k = \begin{pmatrix} \Sigma_{k,1} & 0 & \cdots & 0 \\ 0 & \Sigma_{k,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_{k,H} \end{pmatrix},$$

where the hth block on the diagonal line is defined as

$$\Sigma_{k,h} = \begin{pmatrix} \sigma_{k,h,1,1}^2 & \sigma_{k,h,1,2}^2 & \cdots & \sigma_{k,h,1,g}^2 \\ \sigma_{k,h,2,1}^2 & \sigma_{k,h,2,2}^2 & \cdots & \sigma_{k,h,2,g}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k,h,g,1}^2 & \sigma_{k,h,g,2}^2 & \cdots & \sigma_{k,h,g,g}^2 \end{pmatrix},$$

with $\sigma_{k,h,i,j}^2 = \rho^{|i-j|} \sigma_{k,h,i,i} \sigma_{k,h,j,j}$ and the predefined correlation coefficient ρ. We simulate the diagonal elements $\sigma_{k,h,i,i}^2$ from the uniform distribution, $\mathrm{U}(0.5, 1.5)$. To model the situation with equal covariance matrices between two classes, we set $\Sigma = \Sigma_1 = \Sigma_2$; otherwise, we use $\Sigma_1 \neq \Sigma_2$.

The simulation design of multiple classification is similar to that of a binary classification. For simplicity, we consider the following three-class case where the first g genes are informative with $\mu_{1i} = 0.5, \mu_{2i} = 0$, and $\mu_{3i} = -0.5, i = 1, \ldots, g$. We choose the two-fold cross-validation scheme to estimate CWA. Specifically for each simulation, we randomly take two-thirds of the samples from each class as the training set and the rest as the test set, that is, $n_k/(m_k + n_k) = 2/3$. The average CWA is computed by repeating this random division and testing procedure 100 times for each simulation and then averaging for 1000 simulations.

5.2 *Simulation Results*

Results that assess the CWA change assuming constant variances are shown in Figure 1. The sum of squared mean vector difference is set as $\delta = 10$ (except for the lower right panel of Figure 1). The positive Pr_{Δ} values, that is, overall changes of CWA, indicate that the bias-corrected discriminant rules outperform the original ones (the top panels). The Pr_{Δ} values computed from equation (5) are very close to those from equation (6) even when p is as small as 10. In the upper left panel, we fix the degree of unbalance, $n_2/n_1 = 5$, and vary p. We observe that as p increases, Pr_{Δ} increases sharply first and then decreases slowly after reaching its maximum value. The tail becomes heavier when n_1 increases, that is, the improvement always keeps for large sample sizes. For example, with $n_1 = 20$ and $n_2 = 100$, we still have about 5% gain of CWA at $p = 100$ and the maximum 20.6% is reached at $p = 874$. In the upper right panel, n_2 is fixed at 40. When n_1 changes from 4 to 40, Pr_{Δ} decreases as n_1 increases for small p. For large p, Pr_{Δ} increases first then decreases as n_1

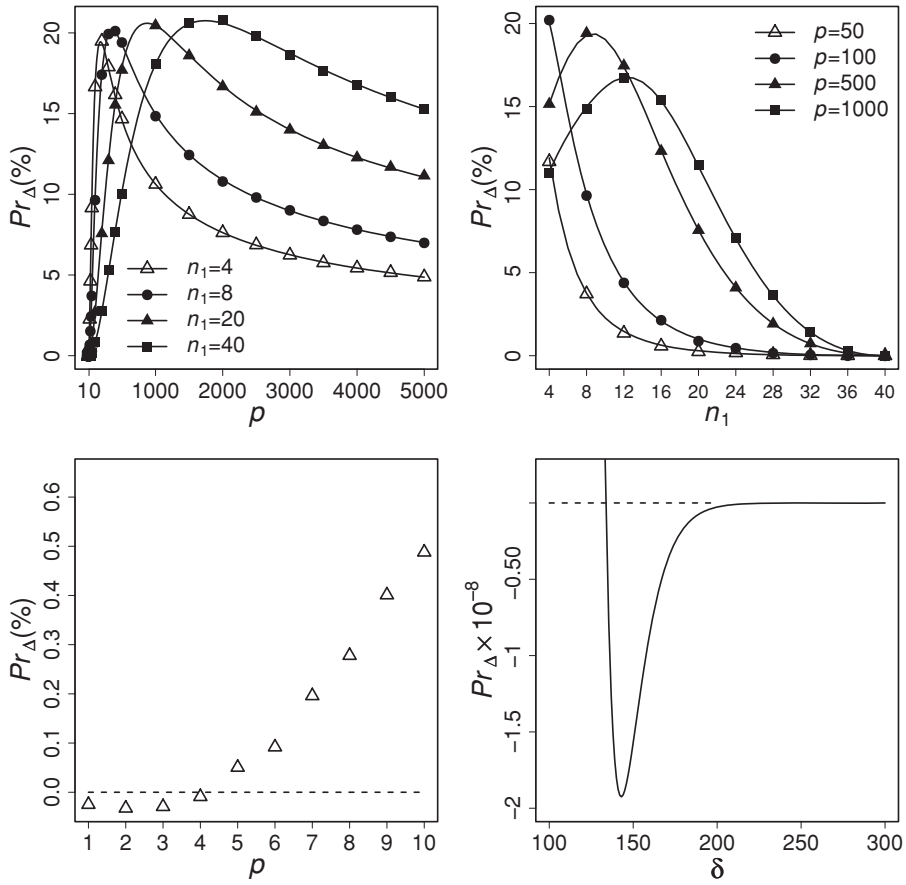

Figure 1. Pr_Δ as functions of different factors (p, n_1, and δ). The solid lines represent results using equation (5). The symbols on the lines represent results using equation (6). The dashed lines represent $Pr_\Delta = 0$. We set $n_2/n_1 = 5$ except for the upper right panel and $\delta = 10$ except for the lower right panel. Upper left: the results for different values of n_1 are shown. Upper right: the results for different values of p are shown ($n_2 = 40$). Lower left: use equation (6) for small $p(n_1 = 4)$. Lower right: use equation (5) for large δ ($n_1 = 4$ and $p = 50$).

increases. For example, when $p = 500$, the maximum improvement of 19.4% can be obtained at $n_1 = 10$. The bottom panels show that we may have $Pr_\Delta < 0$ under certain conditions. The lower left panel shows that Pr_Δ is negative when $p < 5$ and increases with p ($n_1 = 4$ and $n_2 = 20$). The lower right panel shows that Pr_Δ becomes negative when $\delta > 133.9$ and reaches a minimum at $\delta = 142.9 (n_1 = 4, n_2 = 20$, and $p = 50)$.

The bias-corrected discriminant scores do not always outperform the original ones (Section 4). Under certain conditions, we may have $Pr_\Delta < 0$ (Figure 1, the bottom panels). This implies that either (i) a very small number of features is selected, or (ii) a strong signal exists in differentiating the two classes. In practice, a classifier with more than 50 features is often used for classification in microarray analysis (Dudoit et al., 2002; Lee et al., 2005; Golub et al., 1999). Simulation studies suggest Pr_Δ increases rapidly as p increases (Figure 1, the left panels). When the signal is strong, for example, $2p = 100 < \delta$, both bias-corrected methods and the original methods work quite well. Simulation studies suggest that $Pr_\Delta \approx 0$ (Figure 1, the lower right panel).

For the more general simulation settings, we set $\rho = 0.3, H = 200$, and $G = 10{,}000$. We use the equal diagonal covariance matrix to generate samples for both classes. The left column of Figure 2 shows the simulation results for $p = 100, n_2 = 40$, and n_1 varying from 4 to 40. We examine the accuracy of the proposed bias-corrected discriminant scores in terms of the squared biases (Bias^2) and the mean squared errors (MSE) in logarithmic scales for the top and middle panels. We observe that both BQDA and BLDA have smaller Bias^2 and MSE compared with their biased counterparts. When n_1 increases, the difference between unbiased and biased discriminant rules decreases. The bottom panel shows the corresponding results of CWA. Similar to those in Figure 1, the improved prediction accuracy of bias-corrected discriminant scores is consistent; larger improvement happens at smaller sample sizes with higher degrees of unbalance, and becomes indistinguishable for the balanced data.

The right column of Figure 2 displays the effect of p on the estimation and prediction accuracy ($n_1 = 20$ and $n_2 = 100$). It is clear that the bias-corrected discriminant scores provide more accurate and more stable estimates than the original ones consistently (the top and middle panels). The bottom panel shows that the bias-corrected scores have slight improvement when p is small, for example, CWA increases about

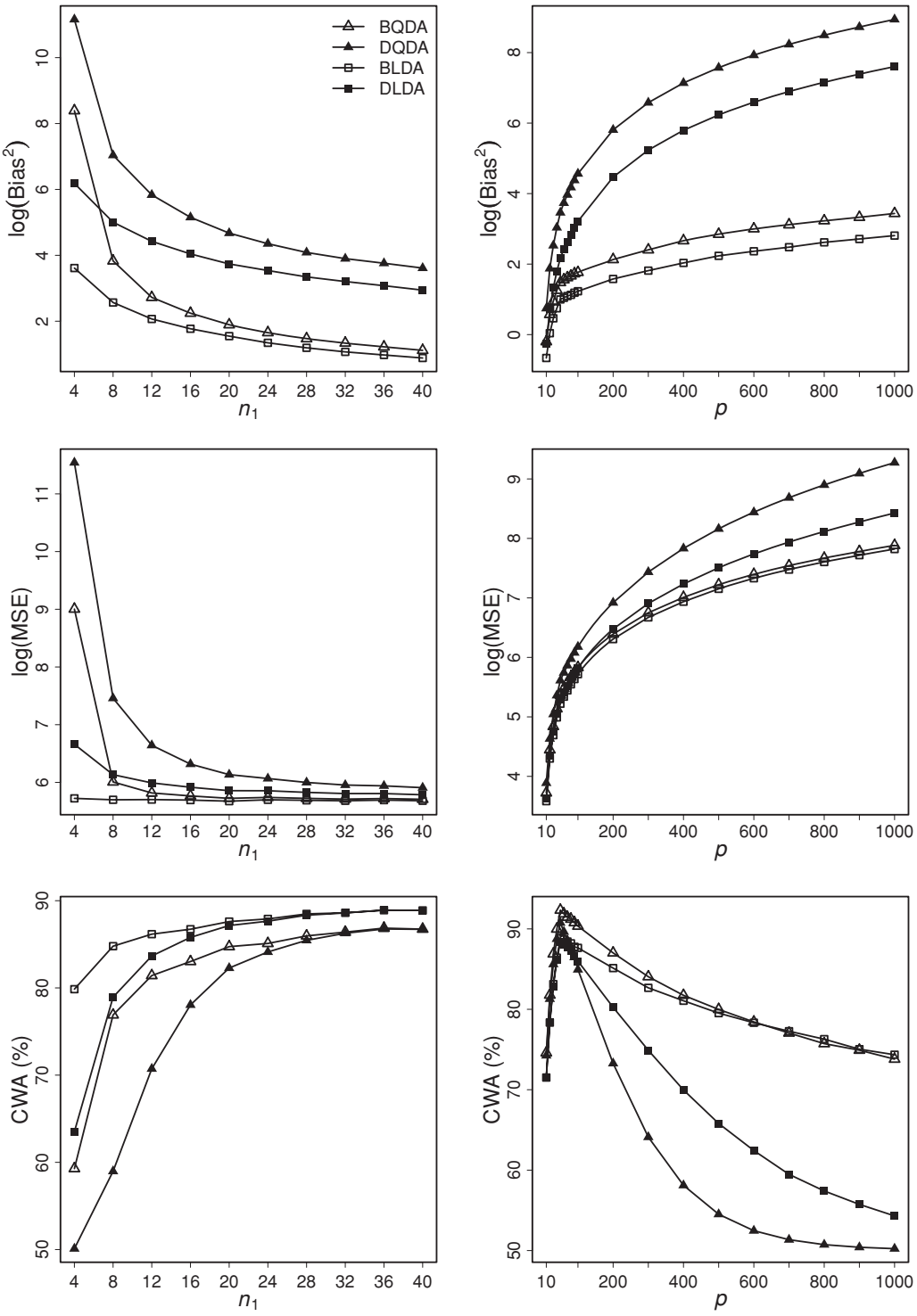

Figure 2. Comparison between bias-corrected discriminant rules and the original ones. Left column: $n_2 = 40$ and $p = 100$. Right column: $n_1 = 20$ and $n_2 = 100$. Top row: $Bias^2$ in a logarithmic scale. Middle row: MSE in a logarithmic scale. Bottom row: CWA.

1% for BQDA versus DQDA when $p = 10$. The improvement becomes more evident when p increases, for example, CWA increases 6.9% for BQDA versus DQDA at $p = 100$. CWA of all methods peak around $p = g$. As more noninformative genes are included in the classifier, the class predication will tend to be random with a final CWA at 50%. However, we observe that even for such situations the bias-corrected discriminant rules still outperform the biased ones.

For multiple classification, we consider three sets of designs: (1) keep $n_1 = n_3$; (2) keep $n_2 = n_3$; (3) keep $n_2/n_3 = 1/2$. For all settings, we vary n_1 and n_2 the same way as in the binary classification settings. In the left columns of Web Figures 1–19, n_1 varies from 4 to 40 with $n_2 = 40$ and $p = 100$. In the right columns, p varies from 10 to 1000 with sample sizes fixed. We observe similar patterns as in binary classification simulation studies. For the results of using unequal covariance matrices and MLE-based discriminant rules (MLDA and MQDA), see Web Figures. As in Guo et al. (2007), we also conduct simulations with a feature selection procedure (Section 6) and incorporate different degrees of correlations, for example, $\rho = 0.5$ or 0.7. The comparisons also have similar patterns as shown for binary classifications (see Web Figures). As the simulation results suggest, the improved performance of the bias-corrected discriminant rules over the original ones is evident for unbalanced classification analyses, especially when the degree of unbalance, for example, the ratio n_2/n_1, is far from 1.

6. Case Studies

In this section, we apply the proposed bias-corrected methods to three real microarray data sets and compare them with several other popular classification methods, including the original diagonal discriminant analysis (DQDA, DLDA, MQDA, and MLDA), support vector machines (SVM), and k-nearest neighbors (kNN). SVM is a supervised machine learning method that aims to find a separating hyperplane into the input space that maximizes the margin between classes (Boser, Guyon, and Vapnik, 1992). It is one commonly used classification method for high-dimensional data with small sample sizes. See for example in Ye et al. (2004), Lee et al. (2005), and Shieh et al. (2006). kNN is a simple algorithm that classifies a sample by the majority voting of its neighbors. This nonparametric classification method is widely used in discriminant analysis and works well in many studies (Dudoit et al., 2002; Lee et al., 2005). In this article, we use the radial basis kernel for SVM and take the three nearest neighbors in Euclidean distance for kNN.

For the binary classification, we first analyze the B-cell lymphoma (BCL) data set in Shipp et al. (2002). The authors applied the weighted voting classification algorithm to differentiate diffuse large B-cell lymphoma (DLBCL) from follicular lymphoma (FL), a related germinal centers BCL. The gene expression data based on oligonucleotide microarray are available for 58 DLBCL and 19 FL pretreatment biopsy samples with 6817 genes. Although DLBCL and FL have different responses to cancer therapy, they share similar morphologic and clinical features over time. The authors showed that the two types of tumors may be distinguished by using their molecular markers. The second data set studied embryonal tumor of central nervous system (CNS), about which little is known biologically, but is believed to have heterogeneous pathogenesis (Pomeroy et al., 2002). The authors investigated the molecular heterogeneity of the most common brain tumor type, medulloblastomas, including primarily the desmoplastic subclass and the classic subclass. The desmoplastic subclass is often seen with a high frequency with Gorlin's syndrome. They analyzed nine desmoplastic samples and 25 classic samples with oligonucleotide microarrays of 6817 genes. The results suggested that the Sonic Hedgehog (SHH) signaling pathway is involved in the pathogenesis of desmoplastic medulloblastoma. In the same study, the authors also investigated the problem of distinguishing multiple types of embryonal CNS tumors at gene expression level. In the original data set, there are 60 medulloblastomas, 10 malignant gliomas, 10 AT/RT (5 CNS, 5 renal-extrarenal), 6 supratentorial PNETs, and 4 normal cerebellums. We exclude the class of normal samples in the study as BQDA requires a minimum of four training samples and one test sample for each class.

All data sets with raw intensity values can be downloaded from the Broad institute website (`http://www.broad.mit.edu`) and are preprocessed with the standard microarray data preprocessing `R` package from Bioconductor (`http://www.bioconductor.org`). We normalize all of the data sets with robust multichip average (RMA) as described in Irizarry et al. (2003). The array control probe sets are removed from analysis after normalization. As in Dudoit et al. (2002), we perform a simple gene selection procedure using the ratios of the between-groups sum of squares (BSS) to the within-groups sum of squares (WSS) for the training set. Specifically, for the jth gene, the ratio is

$$\frac{BSS(j)}{WSS(j)} = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\bar{x}_{k\cdot j} - \bar{x}_{\cdot\cdot j})^2}{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{kij} - \bar{x}_{\cdot\cdot j})^2},$$

where $\bar{x}_{\cdot\cdot j}$ is the averaged expression values across all samples and $\bar{x}_{k\cdot j}$ is that across samples belonging to the kth class. We select the top p genes with the largest BSS/WSS ratios for classification. Similar to the simulation studies in Section 5, we randomly divide the samples of each class into the training set and the test set. The training sample size for the smallest class varies from 4 to $n_1 + m_1 - 1$ (we always set the first class be the smallest one), where $n_1 + m_1$ is the total sample size for the smallest class. For other classes, we hold the same number of samples, m_1, for testing, and use the rest for training. We repeat this procedure 1000 times and report the average CWA for each method.

The results for the binary classification are summarized in Figure 3, where CWA is treated as a function of n_1 with $p = 100$.

It is clear that the performance of the bias-corrected rules is consistently better than that of the original ones. It is also interesting to see that the large improvement by the bias correction may result in the change of order of CWA. For example, in the top right panel, when $n_1 = 4$, BQDA performs the best, even when SVM and kNN have higher CWA than DQDA and MQDA. We observe similar patterns of CWA as p varies (results not shown). Although there is little difference between the MLE and the sample variance estimator when n_k is large, for data sets with small sample sizes, MQDA usually has lower prediction accuracy than DQDA (the top panels) while MLDA performs slightly better than DLDA (the bottom panels).

For the multiclass classification, we show the results with only one test sample but at a series number of selected features p in Table 1. We observe that the bias-corrected

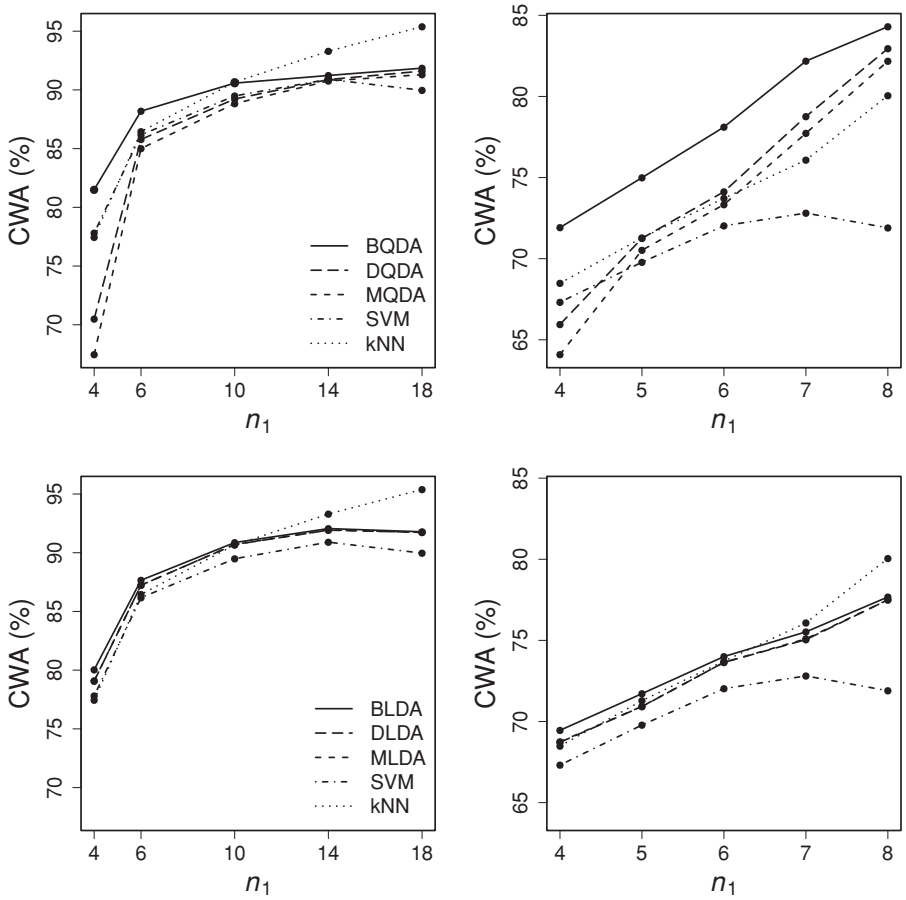

Figure 3. CWA (%) as a function of n_1 for the DLBCL (left) and CNS (right) data sets. The top two panels show the comparison of QDA-based methods; the bottom two panels show the comparison of LDA-based methods. All panels include SVM and *k*NN.

Table 1
CWA (%) *for the multiclass CNS data set*

p	10	50	100	150	200
BQDA	68.480	73.618	**78.438**	**76.160**	**75.180**
DQDA	63.517	68.630	69.735	69.253	69.303
MQDA	62.923	68.152	69.165	68.955	69.185
BLDA	**68.643**	**77.082**	76.040	73.807	72.655
DLDA	68.135	74.703	75.415	73.715	72.137
MLDA	68.230	74.807	75.422	73.723	72.155
SVM	63.415	71.405	74.838	74.038	72.473
*k*NN	61.520	61.830	62.817	63.770	63.570

The top ranked CWA values are in bold text.

discriminant rules outperform the other methods for all p values, among which BQDA performs the best when $p \geqslant 100$ and BLDA performs the best when $p < 100$ (Table 1 with the top ranked CWA highlighted in bold text).

7. Discussion

For high-dimensional data such as microarrays, we face the challenge of building a reliable classifier with a limited number of samples. For instance, a typical microarray study has expression levels for thousands of genes but less than one hundred samples. Much smaller sample sizes (<10) are also common in practice. Diagonal discriminant analysis has been recommended for the high-dimensional data classification problem with remarkably good performance (Dudoit et al., 2002; Lee et al., 2005). However, the conventional estimators of diagonal discriminant scores may not be reliable as they are all biased. In this article, we proposed several bias-corrected discriminant rules that improve the overall prediction accuracy in both simulation studies and real case studies. The bias-corrected methods improve the prediction of the minority class, but sacrifice some performance for the majority class in terms of the per-class prediction accuracy. In reality, the minority class, for example, representing some rare disease samples, is often of interest and may deserve more weight. Here, we show that generally the bias-corrected methods offer higher CWA than the corresponding biased ones, even with the equal weights. The improvement may be affected by many factors, among which the sample size of minority class, n_1, the degree of unbalance, n_1/n_2, and the number of features selected, p, are the most important ones.

When the design is balanced, the bias-corrected rules perform similarly as the original ones, even though the bias-corrected rules provide a better estimator of discriminant scores. Specifically, BLDA performs exactly the same as DLDA, and BQDA performs similarly as DQDA. For unbalanced designs, the change of CWA is nontrivial. As shown

in Sections 5 and 6, the bias-corrected methods outperform their biased counterparts under all simulation settings and real case studies when $n_2/n_1 > 1$ and p is large. The improvement is evident when the sample size of the minority class in the training class (n_1) is small. When n_1 is large, the improvement is still not trivial as long as the ratio of n_2/n_1 keeps. For the DLBCL data set with 18 training samples in the minority class, the overall performance improvement is still observable with only 100 genes selected for classification. To make the bias-corrected rules work, BQDA requires $n_k \geqslant 4$, and BLDA requires $n \geqslant K + 2$ which is less restrictive than BQDA. Most of the publicly available microarray data sets satisfy such requirements.

One possible future research is to propose a regularization between BLDA and BQDA as those in Friedman (1989), Guo et al. (2007), and Pang et al. (2009). To stabilize the variances of $\hat{d}_k^Q(\boldsymbol{y})$ and $\hat{d}_k^L(\boldsymbol{y})$, or equivalently to correct their second-order biases is also of interest. Specifically, we will incorporate the bias correction together with the shrinkage technique in Tong and Wang (2007) and Pang et al. (2009). The rationale behind the shrinkage estimation is to trade off the increased bias for a possible "significant decrease" in the variance (James and Stein, 1961; Radchenko and James, 2008). As a consequence, the good performance of the shrinkage-based discriminant rules is mainly because of the largely reduced variances in the corresponding discriminant scores (Pang et al., 2009). Nevertheless, the bias terms still remain, and more likely, the biases will be larger than that in the original diagonal discriminant scores owing to the impact of shrinkage. Motivated by this, we expect that to correct the biases for the shrinkage-based discriminant rules can be of great interest.

Finally, we reiterate that the diagonal matrix assumption used is somewhat restrictive, so it might be necessary to drop such condition and obtain similar results for more general covariance matrices. Storey and Tibshirani (2001) suggested that the clumpy dependence (i.e., the block diagonal matrix) is a likely form of dependence in the setting of microarray data analysis. This is also mentioned in Langaas, Lindqvist, and Ferkingstad (2005). Inspired by this, one natural extension is to propose the bias-corrected rules for the covariance matrix $\Sigma_k = \mathrm{diag}(\Sigma_{k,1}, \ldots, \Sigma_{k,H})$, where H is the total number of blocks. To calculate the expectation of $\hat{L}_{k1}$ and $\hat{L}_{k2}$ under the block diagonal covariance matrix, the following well-known results can be used (Das Gupta, 1968):

$$E\left(\hat{\Sigma}_k^{-1}\right) = \frac{n_k - 1}{n_k - p - 2} E(\Sigma_k^{-1}),$$

$$E(\log|\hat{\Sigma}_k|) = \log|\Sigma_k| - p\log(n_k - 1) + \sum_{i=1}^{p} \Psi\left(\frac{n_k - 1}{2}\right),$$

where $\hat{\Sigma}_k$ is the sample covariance matrix of class k. In addition, to avoid the singularity problem, it needs to be assumed that each block size is not bigger than the sample size.

8. Supplementary Materials

The Web Appendix and Figures referenced in Sections 3-5 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

The research was supported in part by NIH grant GM59507 and NSF grant DMS0714817. The authors thank Dr Xin Qi for helpful suggestions, and Dr Joshua Sampson for a critical reading and extensive discussions of the article. Part of the simulations were run on the Yale High Performance Computing Cluster, supported by NIH grant RR19895-02. The authors also thank the editor, the associate editor, and two referees for their constructive comments and suggestions that have led to a substantial improvement in the article.

References

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions*. New York: Dover.

Antoniadis, A., Lambert-Lacroix, S., and Leblanc, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics* **19,** 563–570.

Asyali, M. H., Colak, D., Demirkaya, O., and Inan, M. S. (2006). Gene expression profile classification: A review. *Current Bioinformatics* **1,** 55–73.

Bickel, P. J. and Levina, E. (2004). Some theory of Fisher's linear discriminant function, "naive Bayes," and some alternatives when there are many more variables than observations. *Bernoulli* **10,** 989–1010.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–152, Pittsburgh, Pennsylvania.

Braga-Neto, U. M. and Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics* **20,** 374–380.

Breiman, L. (2001). Random forests. *Machine Learning* **45,** 5–32.

Chan, P. (2006). Log-gamma distribution. In *Encyclopedia of Statistical Sciences*, S. Kotz, C. B. Read, N. Balakrishnan, and B. Vidakovic (eds). New York: Wiley.

Chen, C., Liaw, A., and Breiman, L. (2004). Using random forest to learn imbalanced data. Technical Report 666, Department of Statistics, University of California, Berkeley.

Cohen, G., Hilario, M., Sax, H., Hugonnet, S., and Geissbuhler, A. (2006). Learning from imbalanced data in surveillance of nosocomial infection. *Artificial Intelligence in Medicine* **37,** 7–18.

Dabney, A. R. (2005). Classification of microarrays to nearest centroids. *Bioinformatics* **21,** 4148–4154.

Dabney, A. R. and Storey, J. D. (2007). Optimality driven nearest centroid classification from genomic data. *PLoS ONE* **2,** e1002.

Dai, J., Lieu, L., and Rocke, D. (2006). Dimension reduction for classification with gene expression microarray data. *Statistical Applications in Genetics and Molecular Biology* **5,** 6.

Das Gupta, S. (1968). Some aspects of discrimination function coefficients. *Sankhya* **30,** 387–400.

Dettling, M. (2004). Bagboosting for tumor classification with gene expression data. *Bioinformatics* **20,** 3583–3593.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97,** 77–87.

Durrett, R. (1996). *Probability: Theory and Examples*, 2nd edition. Belmont, California: Duxbury Press.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association* **84,** 165–175.

Fu, W., Dougherty, E. R., Mallick, B., and Carroll, R. J. (2005). How many samples are needed to build a classifier: A general sequential approach. *Bioinformatics* **21,** 63–70.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16,** 906–914.

Ghosh, D. (2003). Penalized discriminant methods for the classification of tumors from gene expression data. *Biometrics* **59,** 992–1000.

Ghurye, S. G. and Own, I. (1969). Unbiased estimation of some multivariate probability densities and related functions. *Annals of Mathematical Statistics* **40,** 1261–1271.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286,** 531–537.

Guo, Y., Hastie, T., and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* **8,** 86–100.

Huang, D. and Zheng, C. (2006). Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* **22,** 1855–1862.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of affymetrix genechip probe level data. *Nucleic Acids Research* **31,** e15.

Isaksson, A., Wallman, M., Göransson, H., and Gustafsson, M. G. (2008). Cross-validation and bootstrapping are unreliable in small sample classification. *Pattern Recognition Letters* **29,** 1960–1965.

James, M. (1985). *Classification Algorithms*. New York: Wiley.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **1,** 361–379.

Langaas, M., Lindqvist, B. H., and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society, Series B* **67,** 555–572.

Lee, J. W., Lee, J. B., Park, M., and Song, S. H. (2005). An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics and Data Analysis* **48,** 869–885.

Lee, Y. K., Lin, Y., and Wahba, G. (2004). Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association* **99,** 67–81.

Lehmann, E. L. (1998). *Elements of Large Sample Theory*. New York: Springer.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley.

Moran, M. A. and Murphy, B. J. (1979). A closer look at two alternative methods of statistical discrimination. *Applied Statistics* **28,** 223–232.

Natowicz, R., Incitti, R., Horta, E. G., Charles, B., Guinot, P., Yan, K., Coutant, C., Andre, F., Pusztai, L., and Rouzier, R. (2008). Prediction of the outcome of preoperative chemotherapy in breast cancer using DNA probes that provide information on both complete and incomplete responses. *BMC Bioinformatics* **9,** 149.

Noushath, S., Kumar, G. H., and Shivakumara, P. (2006). Diagonal Fisher linear discriminant analysis for efficient face recognition. *Neurocomputing* **69,** 1711–1716.

Pang, H., Tong, T., and Zhao, H. (2009). Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data. *Biometrics* **65,** 1021–1029.

Pique-Regi, R., Ortega, R., and Asgharzadeh, S. (2005). Sequential diagonal linear discriminant analysis (SeqDLDA) for microarray classification and gene identification. *Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference*, 112–116, Los Alamitos, California.

Pomeroy, S., Tamayo, P., Gaasenbeek, M., Sturla, L., Angelo, M., McLaughlin, M., Kim, J., Goumnerova, L., Black, P., Lau, C., Allen, J., Zagzag, D., Olson, J., Curran, T., Wetmore, C., Biegel, J., Poggio, T., Califano, A., Stolovitzky, G., Louis, D., Mesirov, J., Lander, E., and Golub, T. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415,** 436–442.

Qiao, X. and Liu, Y. (2009). Adaptive weighted learning for unbalanced multicategory classification. *Biometrics* **65,** 159–168.

Radchenko, R. and James, G. M. (2008). Variable inclusion and shrinkage algorithms. *Journal of the American Statistical Association* **103,** 1304–1315.

Shen, R., Ghosh, D., Chinnaiyan, A., and Meng, Z. (2006). Eigengene-based linear discriminant model for tumor classification using gene expression microarray data. *Bioinformatics* **22,** 2635–2642.

Shieh, G., Jiang, Y., and Shih, Y. S. (2006). Comparison of support vector machines to other classifiers using gene expression data. *Communications in Statistics: Simulation and Computation* **35,** 241–256.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., Ray, T. S., Koval, M. A., Last, K. W., Norton, A., Lister, T. A., Mesirov, J., Neuberg, D. S., Lander, E. S., Aster, J. C., and Golub, T. R. (2002). Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine* **8,** 68–74.

Speed, T. P. (2003). *Statistical Analysis of Gene Expression Microarray Data*. London: Chapman and Hall.

Statnikov, A., Wang, L., and Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics* **9,** 319.

Storey, J. D. and Tibshirani, R. (2001). Estimating the positive false discovery rate under dependence, with applications to DNA microarrays. Technical Report 2001–28, Department of Statistics, Stanford University.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences* **99,** 6567–6572.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to (DNA) microarrays. *Statistical Science* **18,** 104–117.

Tong, T. and Wang, Y. (2007). Optimal shrinkage estimation of variances with applications to microarray data analysis. *Journal of the American Statistical Association* **102,** 113–122.

Vapnik, V. and Kotz, S. (2006). *Estimation of Dependences Based on Empirical Data*. New York: Springer.

Wang, S. and Zhu, J. (2007). Improved centroids estimation for the nearest shrunken centroid classifier. *Bioinformatics* **23,** 972–979.

Wu, B. (2006). Differential gene expression detection and sample classification using penalized linear regression models. *Bioinformatics* **22,** 472–476.

Ye, J., Li, T., Xiong, T., and Janardan, R. (2004). Using uncorrelated discriminant analysis for tissue classification with gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **1,** 181–190.

Zhu, J. and Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics* **5,** 427–443.

Received December 2008. *Revised December* 2009.
Accepted December 2009.

Appendix A

Proof of Theorem 1

(i) Recall that $\tilde{d}_k^Q$ is an unbiased estimator of d_k^Q while $\hat{d}_k^Q$ is biased. To verify $\mathrm{var}(\tilde{d}_k^Q) < \mathrm{var}(\hat{d}_k^Q)$, it suffices to show

$$\frac{n_k-2}{n_k-1}\text{var}(\hat{L}_{k1})+\text{cov}(\hat{L}_{k1},\hat{L}_{k2})>0. \qquad \text{(A1)}$$

Denote $J_k=\sum_{i=1}^{p}\frac{(y_i-\mu_{ki})^4}{\sigma_{ki}^4}$, when $n_k>5$, we have

$$\text{var}(\hat{L}_{k1})=\frac{2(n_k-1)^2}{(n_k-3)^2(n_k-5)}\times\left\{J_k+\frac{2(n_k-2)}{n_k}L_{k1}+\frac{(n_k-2)}{n_k^2}p\right\}.$$

For $\text{cov}(\hat{L}_{k1},\hat{L}_{k2})$, note that $\ln\{\hat{\sigma}_{ki}^2(n_k-1)/\sigma_{ki}^2\}\sim\ln\chi^2_{n_k-1}$ (Chan, 2006). We have

$$E\left(\frac{\ln\hat{\sigma}_{ki}^2}{\hat{\sigma}_{ki}^2}\right)=\frac{n_k-1}{(n_k-3)\sigma_{ki}^2}\left\{\Psi\left(\frac{n_k-3}{2}\right)-\ln\left(\frac{n_k-1}{2\sigma_{ki}^2}\right)\right\}.$$

Thus

$$\text{cov}(\hat{L}_{k1},\hat{L}_{k2})=-\frac{2(n_k-1)}{(n_k-3)^2}\left(L_{k1}+\frac{p}{n_k}\right).$$

Then equation (A1) can be simplified to

$$J_k+\frac{n_k^2-3n_k+8}{n_k(n_k-2)}L_{k1}+\frac{n_k+4}{n_k^2(n_k-2)}p>0,$$

which holds for any $n_k>5$.

(ii) The proof of (ii) is skipped since it is essentially the same as that of (i).

Appendix B

Proof of Theorem 2

For ease of notation, denote

$$\nu_t=(-1)^t\delta+U,$$
$$\tau_t^2=4b_t\delta+c.$$

Note that for any integers $0<n_1<n_2$, we have $U=p(\frac{1}{n_1}-\frac{1}{n_2})<2p$. We establish Theorem 2 via the following two steps:

(i) When $0<\delta\leqslant U<2p$ (i.e., $\delta-U\leqslant 0<\delta$). As $\tau_1<\tau_2$, we have

$$\begin{aligned}Pr_{\hat{D},1}&=\Phi\left(\frac{\delta}{\tau_1}\right)-\Phi\left(\frac{\delta-U}{\tau_1}\right)\\&>\Phi\left(\frac{\delta}{\tau_2}\right)-\Phi\left(\frac{\delta-U}{\tau_2}\right)\\&>\Phi\left(\frac{\delta+U}{\tau_2}\right)-\Phi\left(\frac{\delta}{\tau_2}\right)\\&=\Phi\left(\frac{U-\nu_2}{\tau_2}\right)-\Phi\left(\frac{-\nu_2}{\tau_2}\right)\\&=Pr_{\hat{D},2}.\end{aligned}$$

The second inequality is obtained as the standard normal density is a unimodal function and the interval $[\frac{\delta-U}{\tau_2},\frac{\delta}{\tau_2}]$ contains the mode. The last equality is obtained by the symmetry of the standard normal density function.

(ii) When $U<\delta\leqslant 2p$ (i.e., $0<\delta-U<\delta\leqslant 2p$). Denote the length of interval $[\frac{\delta-U}{\tau_1},\frac{\delta}{\tau_1}]$ as $I_1=U/\tau_1$, and the length of interval $[\frac{\delta}{\tau_2},\frac{\delta+U}{\tau_2}]$ as $I_2=U/\tau_2$. We have $I_1>I_2$ as $\tau_1<\tau_2$. Thus by the monotone decreasing property of the $N(0,1)$ density on $(0,\infty)$, as long as the lower bound of I_1 is not larger than that of I_2, that is, if $\frac{\delta-U}{\tau_1}\leqslant\frac{\delta}{\tau_2}$, we can claim that Theorem 2 holds.

In what follows, we verify the condition $\frac{\delta-U}{\tau_1}\leqslant\frac{\delta}{\tau_2}$, or equivalently to verify that

$$\frac{\delta-U}{\delta}\leqslant\frac{\tau_1}{\tau_2}. \qquad \text{(A2)}$$

By the condition that $\delta\leqslant 2p$, the left-hand side of the equation (A2) is

$$\begin{aligned}\text{LHS}&=1-\frac{p}{\delta}\left(\frac{1}{n_1}-\frac{1}{n_2}\right)\\&\leqslant 1-\frac{1}{2}\left(\frac{1}{n_1}-\frac{1}{n_2}\right)\\&=1-\frac{n_2-n_1}{2n_1n_2}.\end{aligned}$$

Meanwhile, by Lemmas 1 and 2 shown below, the right-hand side of equation (A2) is

$$\text{RHS}=\sqrt{\frac{\tau_1^2}{\tau_2^2}}>\sqrt{\frac{4b_1\delta}{4b_2\delta}}=\sqrt{\frac{n_1n_2+n_1}{n_1n_2+n_2}}>1-\frac{n_2-n_1}{2n_1n_2}.$$

Hence, equation (A2) is established and Theorem 2 holds.

Lemma 1. *For any $0<a<b$, the function $f(x)=(a+x)/(b+x)$ is a monotone increasing function of x on $(0,\infty)$.*

Lemma 2. *For any integers $0<n_1<n_2$, we have*

$$\sqrt{\frac{n_1n_2+n_1}{n_1n_2+n_2}}\geqslant 1-\frac{n_2-n_1}{2n_1n_2}.$$

Appendix C

Pr_Δ in equation (6)

Under the assumptions in Section 4, note that if $\boldsymbol{y}$ is given, we can write $\hat{D}=\hat{d}_1^L-\hat{d}_2^L$ as a linear combination of two independent noncentral chi-square random variables, both with p degrees of freedom, i.e.,

$$\hat{D}=\frac{1}{n_1}\chi_p^2(\lambda_1)-\frac{1}{n_2}\chi_p^2(\lambda_2),$$

where $\lambda_k=\sum_{i=1}^{p}n_k(y_i-\mu_{ki})^2$. The expected change of CWA for any fixed observation is defined as $Pr_{\hat{D}|\boldsymbol{y}}=Pr(0<\hat{D}<U|\boldsymbol{y})$. One way to obtain $Pr_{\hat{D}|\boldsymbol{y}}$ is to use the inversion formula of probability characteristic function (Durrett, 1996), that is,

$$Pr_{\hat{D}|\boldsymbol{y}}=\frac{1}{2\pi}\lim_{T\to\infty}\int_{-T}^{T}\frac{1-e^{-i\eta U}}{i\eta}\mathcal{T}(\eta)d\eta,$$

where $\mathcal{T}(\eta)$ is the characteristic function for $\hat{D}$, and

$$\mathcal{T}(\eta)=\left[\frac{n_1n_2}{(n_1-2i\eta)(n_2+2i\eta)}\right]^{\frac{p}{2}}\exp\left(\frac{i\lambda_1\eta}{n_1-2i\eta}-\frac{i\lambda_2\eta}{n_2+2i\eta}\right).$$

To compute Pr_Δ in equation (6), we can sample $\boldsymbol{y}$ from both classes and integrate $Pr_{\hat{D}|\boldsymbol{y}}$ numerically.